基于多种机器学习算法预测广西蔗区甘蔗产量

石杰锋[1]，黄　为[1]，范协洋[1]，李修华[1,2*]，卢阳旭[1]，蒋柱辉[3]，王泽平[4]，
罗　维[1]，张木清[2]

（1. 广西大学 电气工程学院，广西南宁 530004；2. 广西大学甘蔗生物学重点实验室，广西南宁 530004；
3. 广西糖业集团有限公司，广西南宁 530022；4. 广西农业科学院甘蔗研究所，广西南宁 530007）

摘　要：［**目的/意义**］分析广西甘蔗主产区甘蔗产量与气象因素的关系，利用气象数据预测甘蔗产量，为糖厂及相关管理部门提供科学的数据支撑。［**方法**］选用2002～2019年广西五个不同地级市内蔗区的产量数据及14种逐日气象数据，将每年的各气象因子以78个逐月递增的连续时段的均值与产量进行相关性分析，根据敏感时段分析法确定关键气象因子，并分析各气象因子在敏感时段对产量的影响。分别利用BP神经网络（BP Neural Network，BPNN）、支持向量机（Support Vector Machine，SVM）、随机森林（Random Forest，RF）、长短期记忆网络（Long Short-Term Memory，LSTM）建立单蔗区产量预测模型，并采用以全生育期气象均值作为模型输入的方法进行对照实验。使用HP滤波法（Hodrick Prescott Filter）分离出甘蔗气象产量，将5个蔗区的数据混合，分别利用RF、SVM、BPNN和LSTM建立通用的多蔗区气象产量预测模型。［**结果和讨论**］对于单蔗区，敏感时段分析法的模型预测效果明显优于全生育期取气象均值的方法，LSTM模型对于上述两种数据处理方法的预测效果均明显优于目前广泛使用的BPNN、SVM、RF模型，敏感时段分析法的LSTM模型整体的均方根误差（Root Mean Square Error，RMSE）和平均绝对百分比误差（Mean Absolute Percentage Error，MAPE）分别为10.34 t/ha和6.85%，决定系数R_v^2为0.8489。对于多蔗区，LSTM预测结果较差，RF、SVM及BPNN三种预测模型都取得了良好的效果，预测效果最好的BPNN模型的RMSE和MAPE分别为0.98 t/ha和9.59%，R_v^2为0.965。［**结论**］通过敏感时段分析法筛选的关键气象因子与产量均呈显著相关，根据敏感时段能准确地分析各气象因子对产量的影响。使用LSTM模型预测单蔗区产量，使用BPNN模型预测多蔗区甘蔗气象产量的方法是可行的，且预测误差在可接受范围内。

关键词：气象因子；HP滤波；甘蔗产量；BPNN模型；LSTM模型；机器学习

中图分类号：S566.1;S126　　**文献标志码：**A　　**文章编号：**SA202304004

引用格式：石杰锋，黄为，范协洋，李修华，卢阳旭，蒋柱辉，王泽平，罗维，张木清．基于多种机器学习算法预测广西蔗区甘蔗产量[J]. 智慧农业(中英文)，2023，5(2)：82-92.

SHI Jiefeng, HUANG Wei, FAN Xieyang, LI Xiuhua, LU Yangxu, JIANG Zhuhui, WANG Zeping, LUO Wei, ZHANG Muqing. Yield prediction models in Guangxi sugarcane planting regions based on machine learning methods[J]. Smart Agriculture, 2023, 5(2): 82-92. DOI：10.12133/j.smartag.SA202304004 (in Chinese with English abstract)

1　引　言

区域作物产量预测对国家粮食安全评估尤为重要[1]，食糖是国家战略物资，其中87%的产量来自甘蔗，而广西的甘蔗产量位居全国首位，近年来约占全国总产量的70%[2]。对广西甘蔗进行大尺度的估产能为糖厂及相关管理部门提供科学的数据支撑。除土地生产力之外，大田作物产量的高低主要受施肥、灌溉、植保等人为管理因素（生产水平）及降雨、光照、风速等气候因素的制约。地理和人为管理方式通常比较稳定，气候是影响产量的最不可控因素。近年来，越来越多的研究人员投入到基

收稿日期：2023-04-08

基金项目：广西科技重大专项（桂科AA22117004，桂科2018-266-Z01）；国家自然科学基金项目（31760342）

作者简介：石杰锋，研究方向为农业信息化。E-mail：1500807980@qq.com

*通信作者：李修华，博士，副教授，研究方向为作物检测和农业信息化。E-mail：lixh@gxu.edu.cn

于气象数据的农作物产量预测研究中。高俊杰等[3]通过广东省肇庆市高要区多年的气象数据建立早稻的产量预测模型，模型平均准确率为80.23%。于珍珍等[4]使用遗传算法优化神经网络来预测甘蔗产量，R^2达到了0.842。陈上[5]利用陕西杨凌、合阳、长武等地区及各站点的多年历史气象数据建立玉米产量预测模型，整体预测结果趋近于实际产量（平均绝对相对误差＜15%）。基于气象对产量进行预测的研究中，通常需要将产量分割为趋势产量与气象产量。趋势产量通常由地力、管理等因素决定，气象产量由气象决定。目前产量分离方法的研究主要以滑动平均法和Logistic拟合法为主。王二虎和宋晓[6]利用滑动平均的方法将花生气象产量和趋势产量进行分离处理，平均预测精度达到了91%。何虹等[7]利用五点二次平滑法对宁夏引黄灌区玉米的趋势产量与气候产量进行分离，其相对气候产量模型复相关系数均达0.73以上。另外，气象数据因素多、数据量大，通常需要预处理来更有效地提取气象特征。顾雅文等[8]将阿克苏地区苹果进行月平均气象数据与气象产量进行相关性分析，确定敏感月份的气象特征来进行后续建模。何修君[9]将气象日、月、年数据分别与玉米产量进行相关性分析来选取气象特征，分别用三个时间维度数据训练玉米产量预测模型，实现日、月、年三种时间维度的产量预测。李严明[10]将小麦全生育期的气象数据取平均值后与气象产量进行相关性分析，提取得到敏感的气象特征进行建模。针对不同地区的作物产量预测，前人也对多种建模方法进行了尝试。Zhao等[11]使用作物机理和统计回归模型结合的方法来预测小麦产量，相关系数达到了0.86。Croci等[12]利用动静因子及物候的方法发布最佳预测时间，使用高斯过程回归对玉米产量进行预测，最佳性能归一化均方根误差（Normalized Root Mean Square Error，nRMSE）为13.31%。Oikonomidis等[13]使用混合卷积神经网络-深度神经网络（Convolutional Neural Networks-Deep Neural Networks，CNN-DNN）模型在土豆公共数据集上进行预测，模型的预测拟合度为0.87。Di等[14]利用贝叶斯优化的长短时记忆模型预测冬小麦产量，优化后R^2的最好效果为0.80。Burdett和Wellen[15]使用多种机器学习方法分别对玉米和大豆产量进行预测，效果最好的为随机森林（Random Forest，RF）模型，R^2分别为0.85和0.94。

尽管前人在气象产量预测方面进行了大量的研究，但缺乏对于逐月分析气象对甘蔗产量的影响以及适应于多蔗区的产量预测模型的研究。本研究运用广西5个甘蔗主产区内蔗区2002—2019年的日值气象观测资料和甘蔗产量资料，以整年多个连续月份气象均值的数据与产量数据进行相关性分析，确定最优的敏感时段，分析关键气象因子在敏感时段对甘蔗产量的影响。将处理后的数据，利用长短期记忆网络（Long Short-Term Memory，LSTM）与BP神经网络（BP Neural Network，BPNN）、支持向量机（Support Vector Machine， SVM）、RF建立产量预测对比模型，实现单蔗区的产量预测。使用HP滤波法（Hodrick Prescott Filter）分离甘蔗产量，消除不同蔗区之间的差异，将广西5个蔗区数据混合，利用BPNN、SVM、RF建立通用的多蔗区气象产量预测模型，实现多蔗区的产量预测，旨在为广西甘蔗种植管理与相关政策制定提供科学的数据参考。

2 材料与方法

2.1 数据来源

研究区域为位于广西壮族自治区5个不同地级市的蔗区，面积在870～18,500 ha之间。广西蔗区大部分属于亚热带季风气候，年平均气温16.5～23.1℃，大于10℃的积温5000～8300℃，年降水量1300～2000 mm，日照时数1500～1800 h，充足的降雨、日照，以及适宜的气温为甘蔗生长提供了良好的气象条件。各蔗区历年的产量数据由广西糖业集团有限公司（原广西农垦糖业集团）提供，包括2002—2019年共18个榨季的各蔗区总产量（因保密要求不宜公开蔗区具体地级市名称）。各蔗区18年的产量数据统计表如表1所示。

气象数据来自于国家气象科学数据中心（中国气象数据网，http://data.cma.cn）。选取了距离各蔗区直线距离最近的5个气象站点的2002—2019年的日值数据集，包含：20—8时和8—20时降水量

表 1　各蔗区 2002—2019 年产量统计表

Table 1　The statistical information of sugarcane yields of different planting regions from 2002 to 2019

产量/($t \cdot ha^{-1}$)	蔗区 1	蔗区 2	蔗区 3	蔗区 4	蔗区 5
最大值	161.60	91.02	96.72	96.65	126.89
最小值	54.59	58.28	51.75	62.17	37.60
平均值	91.03	72.24	71.01	78.47	66.27

(20 时至第二天 20 时降水量)、极大风速、平均气压、平均 2 min 风速、平均气温、平均水汽压、平均相对湿度、日照时数、最低气压、最低气温、最高气压、最高气温、最大风速、最小相对湿度等 14 个气象因素。各蔗区中心点与对应站点的直线距离均在 40 km 范围内。为提高降水量的空间分辨率，本研究还参考了 Qu 等[16]发表在 Science Data Bank 的 1960～2020 年中国 1 km 分辨率月降水数据集。

2.2 数据预处理

2.2.1　产量数据预处理

为了消除区域性差异，本研究针对多蔗区混合建模时，尝试从甘蔗视在产量分离出由气象因素影响的那部分产量波动，即气象产量。

作物的产量受自然和社会等因素的综合影响。为更好地探究各因素对产量的作用，通常将视在产量分解为趋势产量、气象产量和随机波动产量 3 个分量，如公式（1）所示[17]。

$$Y = Y_t + Y_w + e \tag{1}$$

其中，Y 为视在产量，t/ha；Y_t 为趋势产量，t/ha，由生产水平、土地生产力等因素所决定，具有长期趋势性；Y_w 为气象产量，t/ha，由气象因素所决定；e 为随机噪声。

本研究采用 HP 滤波法[18]分离产量。假设 $\{h_i\}$ ($i=1, 2, \cdots, n$；n 为样本容量）是一个长时间序列，包含长期趋势成分 g_i（本研究指甘蔗的趋势产量）以及短期波动成分 c_i（本研究指甘蔗的气象产量），如公式（2）所示。

$$h_i = g_i + c_i \tag{2}$$

HP 滤波法的主要思想是使长时间序列上的长期趋势成分 g_i 和视在产量序列 h_i 之间偏差的平方和 H 最小，如公式（3）所示。

$$H = \sum_{i=1}^{n} (h_i - g_i)^2 + \lambda \sum_{i=1}^{n} [(g_{i+1} - g_i) - (g_i - g_{i-1})]^2 \tag{3}$$

λ 的取值没有特定的标准，针对不同的时间尺度（如年度、季度、月度等），其取值也有所不同。本研究根据甘蔗年产量数据特点，确定 λ 为 100[19]。

2.2.2　气象数据处理

为了论证本研究气象数据处理方法的科学性，针对单蔗区，使用处理后的气象数据与甘蔗实际产量数据建立产量预测模型，并与前人气象数据处理方法的预测结果进行对照。

甘蔗作物生长期长达 12～14 个月，其产量易受持续性的降雨、大风、低温等气象因素影响，且影响程度因不同生长期而异。前人研究大多以整个生育期内或特定时段的气象均值作为基本因子进行分析，忽略了不同生长期对应的时段长短及主要影响因子具有差异的客观规律，如 1～2 个月的苗期需要适宜的水分，2～3 个月的伸长期需要大量的水分，3～4 个月的成熟期需要控制水分的过量摄入以免影响糖分的积累，生长期后期及成熟期间持续性大风和低温对产量会造成不良影响。

为了寻找最优的时段，本研究采用敏感时段分析法，将甘蔗每年（1—12 月）的气象数据以不同起始月份的按月递增的时间宽度构建了 78 个连续时段（如 1 月开始连续 1 个月、2 个月……，2 月开始连续 1 个月、2 个月……，3 月开始连续 1 个月、2 个月……，以此类推）；然后分别计算不同时段下的气象均值 S_{ij_t}（i，$j=1, 2, \cdots, 12$；$t=1, 2, \cdots, 18$；$i \leqslant j$）。然后针对不同蔗区，分析 18 年来不同时段的 14 种气象均值与甘蔗产量的相关性（公式 4）。

$$r_{ij} = \left| \frac{\sum_{t=1}^{18} (S_{ij_t} - \overline{S_{ij_t}})(Y_t - \overline{Y})}{\sqrt{\sum_{t=1}^{18} (S_{ij_t} - \overline{S_{ij_t}})^2} \sqrt{\sum_{t=1}^{18} (Y_t - \overline{Y})^2}} \right| \tag{4}$$

其中，t 表示年份；i、j 表示月份；S_{ij_t} 为某气象因子第 t 年第 i 月到第 j 月的均值；r_{ij} 为第 i–j 月下某气象因子均值与产量的相关系数，共 78 个。相关系数高说明该因素对甘蔗产量的影响大，可选为模型变量。

2.2.3　数据标准化

气象数据指标多、量纲不一致，会对建模造成一定的影响。另外，本研究除了对单一蔗区进行产量预测，还将基于多个蔗区的混合数据构建通用的

气象产量预测模型；而不同蔗区的产量差异明显，对建模也会造成影响。为了消除气象数据量纲差异及不同蔗区产量差异的影响，本研究将气象数据和产量数据均进行了归一化处理。

常用的归一化方法主要有线性函数归一化和零值归一化。考虑到甘蔗气象产量存在负值，使用线性函数归一化将气象数据与甘蔗产量数据缩放到[-1，1]，如公式（5）所示。

$$X_t = \frac{2\times(X - X_{min})}{X_{max} - X_{min}} - 1 \quad (5)$$

其中，X_t表示归一化后的结果；X表示原始数据；X_{max}、X_{min}分别代表原始数据中的最大、最小值。

2.3 模型构建

本研究采用了在产量预测中广泛使用的RF、SVM及BPNN这3种较传统的算法以及LSTM这一深度学习算法分别建立了单蔗区的视在产量预测模型和多蔗区的气象产量预测模型。

2.3.1 BPNN

BPNN一般由输入层、隐藏层和输出层构成[20]。预测模型的精度主要取决于隐藏层的结构，其层数经多次尝试后确定为1层。模型的输入层为经挑选得到的敏感时段的敏感气象因子作为模型的输入；隐藏层神经元个数通过试凑法确定；输出层节点数为1，即产量。为了避免模型过拟合，将模型训练迭代次数“epoch”设置为200，“dropout”设置为0.3，训练批次大小设置为4。以蔗区1为例，其模型训练过程的损失如图1所示。可以看出，随着迭代次数的增加，训练集和验证集损失不断变小，模型整体不断收敛。

2.3.2 SVM

SVM[21]产量预测模型的构建需要重点关注核函数的选取。本研究对训练数据分别采用线性核函数、多项式核函数与径向基核函数进行建模效果对比，最终选取径向基核函数，该核函数能很好地对不同维度的数据进行非线性映射。其他参数选择了默认值。

2.3.3 RF

RF预测模型的构建需要通过网格交叉搜索的方式遍历参数词典，以寻找最优参数[22]。参数

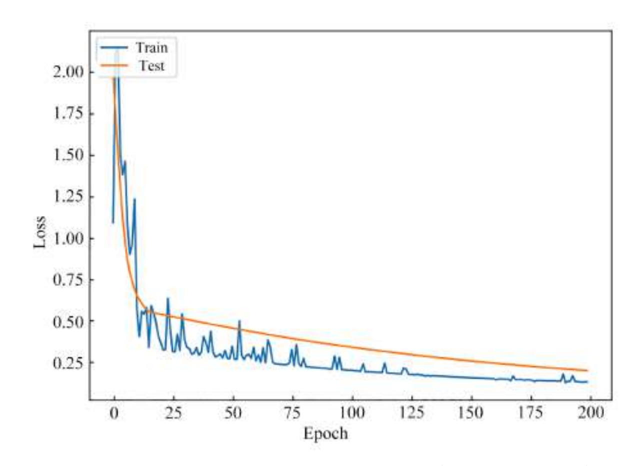


图1 BPNN模型训练过程的损失变化(以蔗区1为例)

Fig.1 The loss changes of the BPNN model (Planting Region 1)

“n_estimators”是每次选取的决策树个数，“max_depth”是RF的最大树深，“max_features”是划分决策树时考虑的最大特征数，各蔗区对应的最优训练参数如表2所示。

表2 各蔗区RF产量预测模型的最优训练参数

Table 2 The optimal training parameters of the BPNN model for each planting region.

最优参数	蔗区1	蔗区2	蔗区3	蔗区4	蔗区5
n_estimators	5	10	5	5	10
max_depth	3	5	5	11	7
max_features	14	14	4	7	4

2.3.4 LSTM

LSTM[23]预测模型含有1个输入层、1个隐含层及1个输出层，其中隐含层拥有25个神经元。模型的时间步长（Time Step）设置为3，训练迭代次数为40，批处理大小为2。

3 结果与讨论

3.1 不同蔗区的关键气象因子分析

3.1.1 气象因子间相关性分析

将每年每个蔗区的14个气象因子在78个不同时段下的均值（共14×78个数据）与产量进行相关性分析。以蔗区1为例，部分具有代表性的气象因子之间以及他们与产量之间的相关系数热力图如图2所示（未显示时段信息）。如果气象因子之间的相关性太高说明自变量间存在较高的自相关，对建模不利；因此需要对相关性较高的多个气象因子

进行筛选，仅保留1个因子作为代表。经相关性分析对因子进行筛选，基本确定了若干个代表性因子，包括日照时数、平均2分钟风速、最大风速、最小相对湿度、平均水气压等。

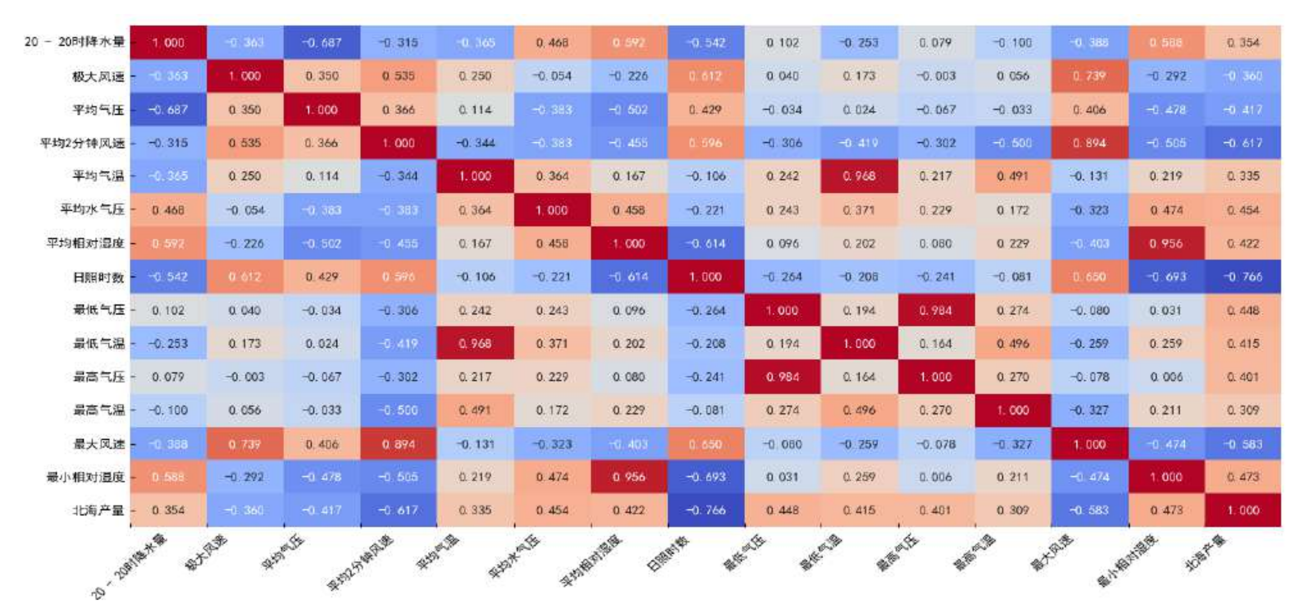


图2　2002—2019年蔗区1气象因子及产量的相关性热力图

Fig. 2 The correlation heat map of different meteorological factors and yields in Planting Region 1 in 2002 to 2019

3.1.2　关键气象因子及敏感时段分析

根据前文的气象数据处理，最终确定与产量关系密切且自相关性低的关键时段的敏感气象因子如表3所示。可以看出，不同蔗区的敏感气象因子以及关键时段均有较大差异。不同时段的气象因子在不同蔗区的相关性甚至相反，这主要是由于甘蔗部分丰欠气象指标跟甘蔗生理气象指标并不完全吻合导致的[24]。

根据相关性结果分析，在分蘖期，日照时数与产量呈显著正相关，良好的光照可以促进分蘖，有利于甘蔗增产。在伸长期后期至成熟期，日照时数与产量呈显著负相关，该时期降水量较少，过于充足的光照容易导致甘蔗枯死，阻碍甘蔗增产。在幼苗期至分蘖期，甘蔗产量与最大风速呈显著负相关，甘蔗幼苗受最大风速影响容易发生折断，折断后严重影响后期生长而导致产量降低。

在萌芽期和伸长期，平均水汽压与甘蔗产量呈显著正相关。平均水汽压与降水量及各种温度、湿度均呈显著正相关，主要通过影响其他气象因素而影响甘蔗产量，是一个综合性气象指标，因此利用平均水汽压评估气象条件的好坏及粗略预估甘蔗产量的发展趋势具有一定意义。

在萌芽期，最低气温与产量呈显著正相关，适宜的最低气温有利于种茎内酶活性的提高，加快萌芽。在幼苗期，平均气温与产量呈显著正相关，气温适当升高有利于幼苗生长。在伸长期和成熟期，最高气温、平均气温均与产量呈显著负相关，该时期光照充足，持续高温加剧了甘蔗枯死，使甘蔗减产。在伸长期，最低气温与产量呈显著负相关，该时期气温日较差小，不利于甘蔗光合作用及糖分的积累，导致减产，因此在该伸长期降低夜间环境温度有利于增产。

多种气象因素与各种气压之间呈显著正相关，但与产量呈显著负相关，主要影响时段为幼苗期，而气压对幼苗生长的作用是间接的，主要通过影响其他气象因子起作用，高压下不利于幼苗生长。

在伸长期，降水量与产量呈极显著负相关，甘蔗主要通过根部吸收水分，该时段降水量过多会导致水淹蔗田，甘蔗根部长期缺氧会导致烂根死亡，因此降雨量较多应及时将水排出，防止甘蔗减产。在伸长期至成熟期，降水量与产量呈显著正相关，甘蔗在这两个时期对水的需求量很大，进一步说明该时期降水量不满足甘蔗生长的水分需求，增加该时段的降水量有利于甘蔗增产。在成熟期后期，最

表3 各蔗区筛选得到的敏感气象因子和关键时段

Table 3 The sensitive meteorological factors and key time spans found for the five planting regions

蔗区	气象因子	与产量的相关系数	影响时段
蔗区1	日照时数	-0.766	10—11月
	平均2分钟风速	-0.617	10—11月
	最大风速	-0.583	4—5月
	最小相对湿度	0.473	11月
	平均水气压	0.454	6月
蔗区2	平均水气压	0.663	2—3月
	最低气温	0.648	2—3月
	最低气压	-0.606	3月
	平均气温	0.596	2—4月
	20时至第二天20时降水量	0.527	8月
蔗区3	20时至第二天20时降水量	0.776	8—9月
	日照时数	-0.555	8—9月
	最低气温	0.542	3月
	平均水气压	0.487	3月
	最高气温	-0.465	8—11月
蔗区4	最高气温	-0.672	8—12月
	平均气温	-0.570	7—12月
	20时至第二天20时降水量	0.657	3—12月
	日照时数	0.502	5月
	最低气温	-0.448	8月
蔗区5	平均水气压	0.829	6—8月
	平均相对湿度	0.715	2—10月
	最低气压	-0.697	2—3月
	最高气压	-0.696	2—3月
	20时至第二天20时降水量	-0.437	6—7月

小相对湿度与甘蔗产量呈显著正相关，该时期干旱也会导致甘蔗减产。平均相对湿度对产量的影响几乎贯穿甘蔗的整个生育期，与产量呈显著正相关，平均相对湿度主要受降雨量影响，增大降雨量，有利于甘蔗增产。由此可知，甘蔗在全生育期对水的需求量很大，但在某一时期降水量过多或过少都不利于甘蔗增产，因此适宜的降水对甘蔗生长尤为重要。

3.2 单蔗区实际产量建模

将上述筛选得到的各蔗区关键时段的敏感气象因子和产量数据标准化，然后将2002—2019年共18年的甘蔗产量数据按7∶3的比例随机划分为训练集与测试集，再分别采用BPNN、SVM、RF以及LSTM方法构建各蔗区的视在产量预测模型，并进行对比。为反映本研究提出的敏感气象因子分析方法的有效性，特采用同样的模型算法基于李严明[10]所用的年均气象因子构建产量模型，并对比二者结果。

3.2.1 BPNN预测模型

各蔗区的BPNN视在产量模型的预测结果如表4所示。基于本文分析所得关键时段敏感气象因子建立的BPNN模型整体上均明显优于对照方法。在5个蔗区中，蔗区4的预测结果最好，产量预测值与实际值之间的均方根误差（Root Mean Square Error，RMSE）仅为5.04 t/ha，平均绝对百分比误差（Mean Absolute Percentage Error，MAPE）仅为5.62%；预测结果最差的为蔗区1，RMSE高达为24.34 t/ha。

表4 各蔗区BPNN模型的视在产量预测结果

Table 4 The apparent yield prediction results of the BPNN models for each planting region

蔗区	本研究方法		对照方法	
	RMSE/($t \cdot ha^{-1}$)	MAPE/%	RMSE/($t \cdot ha^{-1}$)	MAPE/%
蔗区1	24.34	11.88	37.125	32.29
蔗区2	9.25	9.37	14.145	18.19
蔗区3	8.51	10.05	23.64	36.80
蔗区4	5.04	5.62	9.02	9.71
蔗区5	11.37	18.08	17.67	29.75

3.2.2 SVM预测模型

各蔗区的SVM视在产量模型的预测结果如表5所示。可以看出，基于本研究所提数据处理方法建立的SVM模型同样均明显优于对照方法。蔗区2和蔗区4的MAPE均在10%以内，整体上比BPNN模型精度略小，但效果提升不明显，且对蔗区1、蔗区5仍存在较大的预测误差。

3.2.3 RF预测模型

各蔗区的RF视在产量模型的预测结果如表6

表5 各蔗区SVM模型的视在产量预测结果

Table 5 The apparent yield prediction results of the SVM models for each planting region

蔗区	本研究方法		对照方法	
	RMSE/(t·ha⁻¹)	MAPE/%	RMSE/(t·ha⁻¹)	MAPE/%
蔗区1	17.94	8.76	33.20	34.57
蔗区2	9.71	10.34	16.88	16.62
蔗区3	11.01	14.13	18.71	22.27
蔗区4	9.47	6.92	18.36	14.14
蔗区5	11.98	19.99	16.86	24.92

所示。由本研究所提气象因子分析方法构建的RF模型中，除了蔗区1和蔗区5外，其他蔗区均有较高精度，RMSE最低为3.5 t/ha，最高为7.33 t/ha，MAPE在3.98%~7.48%之间。RF模型整体优于SVM和BPNN模型，然而对蔗区1和蔗区5还是存在较大的预测误差，RMSE分别达到了31.83 t/ha和11.22 t/ha。与对照的数据处理方法相比，本研究方法同样具有显著优势。

表6 各蔗区RF模型的视在产量预测结果

Table 6 The apparent yield prediction results of the RF models for each planting region

蔗区	本文方法		对照方法	
	RMSE/(t·ha⁻¹)	MAPE/%	RMSE/(t·ha⁻¹)	MAPE/%
蔗区1	31.83	13.03	39.62	37.85
蔗区2	7.33	7.48	11.52	14.77
蔗区3	5.41	6.20	18.02	21.56
蔗区4	3.50	3.98	9.18	10.61
蔗区5	11.22	19.16	29.64	34.74

3.2.4 LSTM预测模型

LSTM模型的视在产量预测结果如表7所示。由本研究所提气象因子分析方法构建的LSTM模型在5个蔗区均取得较好的结果，各蔗区的MAPE均在10%以内，整体精度相比于其他3种模型有较大提高。与对照方法相比，预测精度也有显著的提高。对照数据处理方法所训练的LSTM模型预测精度相比于其他三种模型精度也有明显提高。可以看出，采用LSTM模型预测甘蔗产量的方法是最优的。

3.2.5 模型预测结果对比分析

将基于同一种模型算法建立的各蔗区视在产量预测模型的测试集结果汇总，展示其预测值与实际值的散点图（图3），以评价各算法在不同蔗区上的整体预测效果。

表7 各蔗区LSTM模型的实际产量预测结果

Table 7 The overall yield prediction results of the LSTM models for each planting region

蔗区	本文方法		对照方法	
	RMSE/(t·ha⁻¹)	MAPE/%	RMSE/(t·ha⁻¹)	MAPE/%
蔗区1	19.96	8.78	28.15	22.72
蔗区2	8.14	7.50	13.59	15.85
蔗区3	5.96	6.23	17.73	27.70
蔗区4	2.60	2.35	9.29	10.68
蔗区5	5.22	9.40	15.23	17.90

基于不同气象敏感因子提取方法的各模型在所有蔗区中的视在产量预测结果如表8所示。可以看出，经本文方法构建LSTM模型，其测试集的整体RMSE和MAPE最低，分别为10.34 t/ha和6.85%，甘蔗产量实际值与预测值之间的决定系数R_v^2最高，达到了0.8489，产量趋势比其他三个模型更为拟合。根据表1的各蔗区甘蔗产量统计数据可知，使用LSTM模型预测甘蔗产量的方法其误差在可接受范围内。使用年均气象因子[10]构建的参照模型的预测精度明显更低。说明本文所提数据处理方法能明显提高甘蔗产量预测模型的精度。由对比结果可知，LSTM产量预测模型预测效果优于BPNN模型、SVM模型和RF模型。

3.3 多蔗区气象产量建模

为了实现大区域甘蔗产量预测，本研究使用HP滤波[25]分离得到各蔗区每年的气象产量，再将5个蔗区的数据混合，建立通用的气象产量预测模型。建模方法主要采用了BPNN、SVM、RF和LSTM，其结果如表9所示。可以看出，LSTM模型在多蔗区中的预测结果较差，其余三种预测模型都取得了良好的效果，R_v^2均在0.94以上；BPNN模型的效果最好，RMSE和MAPE最低，分别为0.98 t/ha和9.59%，R_v^2最高，达到了0.965。

本研究基于5个蔗区数据建立的多蔗区气象产量预测模型，采样了混合数据的方法，有效地扩充了数据集，明显地提升了模型预测效果。考虑到区

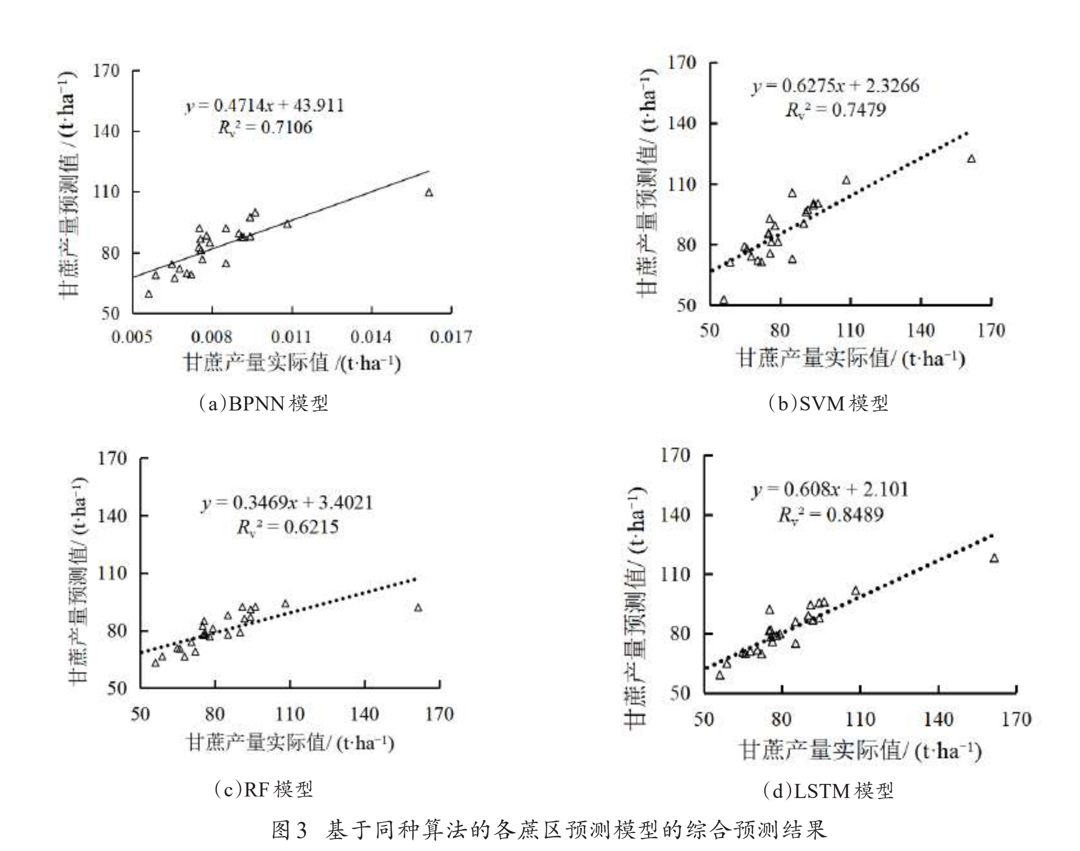


图 3 基于同种算法的各蔗区预测模型的综合预测结果

Fig. 3 The comprehensive prediction results of all the regional-specific models based on same algorithm

表 8 不同模型甘蔗产量预测结果对比

Table 8 Yield prediction result comparison between different models

模型	本文方法			对照方法		
	RMSE/(t·ha^{-1})	MAPE/%	R_v^2	RMSE/(t·ha^{-1})	MAPE/%	R_v^2
BPNN	13.45	11.01	0.7106	22.50	21.15	0.4565
SVM	12.41	12.03	0.7479	22.02	22.51	0.3425
RF	15.71	9.97	0.6215	24.45	23.90	0.3937
LSTM	10.34	6.85	0.8489	11.99	14.63	0.5099

表 9 不同模型的气象产量预测结果对比

Table 9 Meteorological yield prediction results comparison between different models

模型	RMSE/(t·ha^{-1})	MAPE/%	R_v^2
SVM	1.29	13.29	0.944
RF	1.04	9.96	0.957
BPNN	0.98	9.59	0.965
LSTM	0.25	39.99	0.770

域性差异的影响，通过HP滤波分离出甘蔗气象产量，消除了由于不同蔗区的生产条件及社会经济等因素引起的差异，由此建立的通用模型更科学。本研究对气象数据的处理也具有科学性与独特性，由对比结果可知，基于计算后的气象数据与产量数据建立的通用模型，对多区域大范围甘蔗产量的预测是可行的。

4 结 论

根据敏感时段分析法，对78个以月为单位的连续时段气象均值数据与产量进行相关性分析，得到广西5个蔗区的关键气象因子为日照时数、平均水气压、气压、温度、降水量；不同气象因子的关键时段不同，同一气象因子在不同时段与产量的相关性甚至相反，因此根据敏感时段分析关键气象因子对产量的影响具有现实意义。由于不同模型的原理和特点不同，对不同数据集的表现不同，因此不

同甘蔗产量模型的预测结果存在差异。

单蔗区预测模型结果表明，基于本研究分析所得关键时段敏感气象因子的BPNN、SVM、RF以及LSTM四种模型的预测效果均明显优于参考文献[10]的对照方法。LSTM模型的整体RMSE和MAPE分别为10.34 t/ha和6.85%，R_v^2为0.8489，预测效果要明显优于前人研究使用较多的其他三种模型。BPNN、SVM和RF三种预测模型，整体精度较高的模型在部分蔗区的预测精度可能较低，但LSTM模型不仅整体预测效果最好，且对各个蔗区的MAPE均低于10%，因此使用LSTM模型预测各蔗区的产量是可行的。

针对多蔗区，分别使用SVM、RF、BPNN以及LSTM四种模型预测混合样本的甘蔗气象产量，实现多蔗区气象产量的预测。结果表明，BPNN模型整体的RMSE和MAPE分别为0.98 t/ha、9.59%，R_v^2为0.965，其预测效果优于SVM、RF以及LSTM模型，但除LSTM外其余三种模型都取得了良好的效果。LSTM不适用于多蔗区甘蔗产量联合预测。

由此可见，经过本研究方法建立通用的多蔗区气象产量预测模型是可行的。对于单蔗区和多蔗区的产量预测误差均在可接受范围内，本研究预测方法对区域内甘蔗产量预测具有一定的参考意义。

利益冲突声明：本研究不存在研究者以及与公开研究成果有关的利益冲突。

参考文献：

[1] 李威, 顾峰雪. 区域作物产量的模型预测研究[J]. 农业展望, 2020, 16(3): 104-111.
LI W, GU F X. Prediction of regional crop yield based on model[J]. Agricultural outlook, 2020, 16(3): 104-111.
[2] 农业农村部市场预警专家委员会. 中国农业展望报告2019—2028[M]. 北京: 中国农业科学技术出版社, 2019.
Expert Committee on Market Warning of Ministry of Agriculture and Rural Affairs. China agricultural outlook 2019－2028[M]. Beijing: China Agricultural Science and Technology Press, 2019.
[3] 高俊杰, 袁业溶, 梁应. 高要区早稻产量预测模型的建立[J]. 广东气象, 2022, 44(2): 50-52.
GAO J J, YUAN Y R, LIANG Y. Establishment of early rice yield prediction model in Gaoyao area[J]. Guangdong meteorology, 2022, 44(2): 50-52.
[4] 于珍珍, 邹华芬, 于德水, 等. 融合田间水热因子的甘蔗产量GA-BP预测模型[J]. 农业机械学报, 2022, 53(10): 277-283.
YU Z Z, ZOU H F, YU D S, et al. Sugarcane yield GA-BP prediction model incorporating field water and heat factors[J]. Transactions of the Chinese society for agricultural machinery, 2022, 53(10): 277-283.
[5] 陈上. 基于历史气象数据和*CERES*-maize模型的玉米产量预测及灌溉决策方法[D]. 杨凌: 西北农林科技大学, 2017.
CHEN S. Yield forecast and irrigation decision for maize based on historical weather data and the *Ceres*-maize model[D]. Yangling: Northwest A & F University, 2017.
[6] 王二虎, 宋晓. 基于气象因子的开封市花生产量预测模型[J]. 陕西农业科学, 2012, 58(4): 31-33.
WANG E H, SONG X. Prediction model of peanut yield in Kaifeng city based on meteorological factors[J]. Shaanxi journal of agricultural sciences, 2012, 58(4): 31-33.
[7] 何虹, 王巧娟, 李亮, 等. 宁夏引黄灌区玉米趋势产量与气候产量分离方法研究[J]. 灌溉排水学报, 2022, 41(4): 30-39.
HE H, WANG Q J, LI L, et al. Separating the effect of meteorology on maize yield from the impact of other factors in the Yellow River-water irrigated regions in Ningxia of China[J]. Journal of irrigation and drainage, 2022, 41(4): 30-39.
[8] 顾雅文, 姚艳丽, 傅玮东. 基于关键气象因子的阿克苏地区苹果产量预测模型[J]. 新疆农业科技, 2021(2): 22-24.
GU Y W, YAO Y L, FU W D. Prediction model of apple yield in Aksu region based on key meteorological factors[J]. Xinjiang agricultural science and technology, 2021 (2): 22-24.
[9] 何修君. 基于机器学习的玉米产量预测模型研究[D]. 长春: 吉林农业大学, 2021.
HE X J. Research on maize yield prediction model based on machine learning[D]. Changchun: Jilin Agricultural University, 2021.
[10] 李严明. 基于机器学习的气象因素对小麦产量影响的分析预测[D]. 郑州: 河南农业大学, 2019.
LI Y M. Wheat yield forecasting: A machine learning approach based on meteorological factors[D]. Zhengzhou: Henan Agricultural University, 2019.
[11] ZHAO Y X, XIAO D P, BAI H Z, et al. The prediction of wheat yield in the North China plain by coupling crop model with machine learning algorithms[J]. Agriculture, 2022, 13(1): ID 99.
[12] CROCI M, IMPOLLONIA G, MERONI M, et al. Dynamic maize yield predictions using machine learning on multi-source data[J]. Remote sensing, 2022, 15(1): ID 100.
[13] OIKONOMIDIS A, CATAL C, KASSAHUN A. Hybrid deep learning-based models for crop yield prediction[J].

Applied artificial intelligence, 2022, 36(1): 1-18.

[14] DI Y, GAO M F, FENG F K, et al. A new framework for winter wheat yield prediction integrating deep learning and Bayesian optimization[J]. Agronomy, 2022, 12(12): ID 3194.

[15] BURDETT H, WELLEN C. Statistical and machine learning methods for crop yield prediction in the context of precision agriculture[J]. Precision agriculture, 2022, 23(5): 1553-1574.

[16] QU L S, ZHU Q A, ZHU C F, et al. 2022. Monthly precipitation data set with 1 km resolution in China from 1960 to 2020[DB/OL]. Science Data Bank. [2022-04-15]. http://www.scidb.cn/cstr/31253.11.sciencedb.01607.

[17] 黄海迅, 周筠珺, 曾勇, 等. 广西贵港甘蔗产量气象预报[J]. 成都信息工程大学学报, 2020, 35(5): 554-559.
HUANG H X, ZHOU Y J, ZENG Y, et al. Meteorological forecast of sugarcane production in Guigang, Guangxi[J]. Journal of Chengdu university of information technology, 2020, 35(5): 554-559.

[18] 许鑫, 马兆务, 熊淑萍, 等. 基于气候年型的河南省冬小麦产量预测[J]. 中国农业科技导报, 2022, 24(2): 136-144.
XU X, MA Z W, XIONG S P, et al. Wheat yield forecast in Henan Province based on climate year type[J]. Journal of agricultural science and technology, 2022, 24(2): 136-144.

[19] 王桂芝, 陆金帅, 陈克垚, 等. 基于HP滤波的气候产量分离方法探讨[J]. 中国农业气象, 2014, 35(2): 195-199.
WANG G Z, LU J S, CHEN K Y, et al. Exploration of method in separating climatic output based on HP filter[J]. Chinese journal of agrometeorology, 2014, 35(2): 195-199.

[20] ZHOU C H, WU Z Y, LIU C. A study on quality prediction for smart manufacturing based on the optimized BP-AdaBoost model[C]// 2019 IEEE International Conference on Smart Manufacturing, Industrial & Logistics Engineering (SMILE). Piscataway, NJ, USA: IEEE, 2020: 1-3.

[21] KAZEMI A, BOOSTANI R, ODEH M, et al. Two-layer SVM, towards deep statistical learning[C]// 2022 International Engineering Conference on Electrical, Energy, and Artificial Intelligence (EICEEAI). Piscataway, NJ, USA: IEEE, 2022.

[22] MIAH M O, KHAN S S, SHATABDA S, et al. Improving detection accuracy for imbalanced network intrusion classification using cluster-based under-sampling with random forests[C]// 2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT). Piscataway, NJ, USA: IEEE, 2019: 1-5.

[23] AKANDEH A, SALEM F M. Slim LSTM networks: Lstm_6 and LSTM_C6[C]// 2019 IEEE 62nd International Midwest Symposium on Circuits and Systems (MWSCAS). Piscataway, NJ, USA: IEEE, 2020: 630-633.

[24] 欧钊荣, 谭宗琨, 何燕, 等. 影响我国甘蔗主产区甘蔗产量的关键气象因子及其丰欠指标[J]. 安徽农业科学, 2008, 36(24): 10407-10410, 10415.
OU Z R, TAN Z K, HE Y, et al. The key meteorological factors affecting the sugarcane yield in major production areas in China and their high-low yield indices[J]. Journal of Anhui agricultural sciences, 2008, 36(24): 10407-10410, 10415.

[25] 李志强, 张香燕, 田华东. 应用HP滤波的卫星遥测数据预测方法[J]. 航天器工程, 2021, 30(4): 23-30.
LI Z Q, ZHANG X Y, TIAN H D. Prediction method of satellite telemetry data using HP filter[J]. Spacecraft engineering, 2021, 30(4): 23-30.

Yield Prediction Models in Guangxi Sugarcane Planting Regions Based on Machine Learning Methods

SHI Jiefeng[1], HUANG Wei[1], FAN Xieyang[1], LI Xiuhua[1,2*], LU Yangxu[1], JIANG Zhuhui[3], WANG Zeping[4], LUO Wei[1], ZHANG Muqing[2]

（1. *School of Electrical Engineering*, *Guangxi University*, *Nanning* 530004, *China*; 2. *Guangxi Key Laboratory of Sugarcane Biology*, *Guangxi University*, *Nanning* 530004, *China*; 3. *Guangxi Sugar Industry Group*, *Nanning* 530022, *China*; 4. *Sugarcane Research Institute*, *Guangxi Academy of Agricultural Sciences*, *Nanning* 530007, *China*）

Abstract:

[Objective] Accurate prediction of changes in sugarcane yield in Guangxi can provide important reference for the formulation of rele-

vant policies by the government and provide decision-making basis for farmers to guide sugarcane planting, thereby improving sugarcane yield and quality and promoting the development of the sugarcane industry. This research was conducted to provide scientific data support for sugar factories and related management departments, explore the relationship between sugarcane yield and meteorological factors in the main sugarcane producing areas of Guangxi Zhuang Autonomous Region.

[Methods] The study area included five sugarcane planting regions which laid in five different counties in Guangxi, China. The average yields per hectare of each planting regions were provided by Guangxi Sugar Industry Group which controls the sugar refineries of each planting region. The daily meteorological data including 14 meteorological factors from 2002 to 2019 were acquired from National Data Center for Meteorological Sciences to analyze their influences placed on sugarcane yield. Since meteorological factors could pose different influences on sugarcane growth during different time spans, a new kind of factor which includes meteorological factors and time spans was defined, such as the average precipitation in August, the average temperature from February to April, etc. And then the inter-correlation of all the meteorological factors of different time spans and their correlations with yields were analyzed to screen out the key meteorological factors of sensitive time spans. After that, four algorithms of BP neural network (BPNN), support vector machine (SVM), random forest (RF), and long short-term memory (LSTM) were employed to establish sugarcane apparent yield prediction models for each planting region. Their corresponding reference models based on the annual meteorological factors were also built. Additionally, the meteorological yields of every planting region were extracted by HP filtering, and a general meteorological yield prediction model was built based on the data of all the five planting regions by using RF, SVM BPNN, and LSTM, respectively.

[Results and Discussions] The correlation analysis showed that different planting regions have different sensitive meteorological factors and key time spans. The highly representative meteorological factors mainly included sunshine hours, precipitation, and atmospheric pressure. According to the results of correlation analysis, in Region 1, the highest negative correlation coefficient with yield was observed at the sunshine hours during October and November, while the highest positive correlation coefficient was found at the minimum relative humidity in November. In Region 2, the maximum positive correlation coefficient with yield was observed at the average vapor pressure during February and March, whereas the maximum negative correlation coefficient was associated with the precipitation in August and September. In Region 3, the maximum positive correlation coefficient with yield was found at the 20–20 precipitation during August and September, while the maximum negative correlation coefficient was related to sunshine hours in the same period. In Region 4, the maximum positive correlation coefficient with yield was observed at the 20–20 precipitation from March to December, whereas the maximum negative correlation coefficient was associated with the highest atmospheric pressure from August to December. In Region 5, the maximum positive correlation coefficient with yield was found at the average vapor pressure from June and to August, whereas the maximum negative correlation coefficient as related to the lowest atmospheric pressure in February and March. For each specific planting region, the accuracy of apparent yield prediction model based on sensitive meteorological factors during key time spans was obviously better than that based on the annual average meteorological values. The LSTM model performed significantly better than the widely used classic BPNN, SVM, and RF models for both kinds of meteorological factors (under sensitive time spans or annually). The overall root mean square error (RMSE) and mean absolute percentage error (MAPE) of the LSTM model under key time spans were 10.34 t/ha and 6.85%, respectively, with a coefficient of determination R_v^2 of 0.8489 between the predicted values and true values. For the general prediction models of the meteorological yield to multiple the sugarcane planting regions, the RF, SVM, and BPNN models achieved good results, and the best prediction performance went to BPNN model, with an RMSE of 0.98 t/ha, MAPE of 9.59%, and R_v^2 of 0.965. The RMSE and MAPE of the LSTM model were 0.25 t/ha and 39.99%, respectively, and the R_v^2 was 0.77.

[Conclusions] Sensitive meteorological factors under key time spans were found to be more significantly correlated with the yields than the annual average meteorological factors. LSTM model shows better performances on apparent yield prediction for specific planting region than the classic BPNN, SVM, and RF models, but BPNN model showed better results than other models in predicting meteorological yield over multiple sugarcane planting regions.

Key words: meteorological factor; HP filter; sugarcane yield; BPNN model; LSTM model; machine learning